Journal of Image and Signal Processing 图像与信号处理, 2013, 2, 37-41
http://dx.doi.org/10.12677/jisp.2013.23006 Published Online July 2013 (http://www.hanspub.org/journal/jisp.html)

Mathematical Formula Automatic Location Method Based on Circular Projection Statistics*

Xiaoyang Peng[1], Jianpin Mao[2]

[1]College of Economics and Management, Shaoyang University, Shaoyang
[2]Fuzhou Vocational and Technical College, Fuzhou
Email: ppakaka@qq.com

Received: Apr. 13th, 2013; revised: Apr. 21st, 2013; accepted: May 20th, 2013

Abstract: The location of mathematical formulas is the first step to recognize mathematical formula. Only when the formula in the document image is located correctly, one can complete the following steps such as formula symbol recognition, formula document analysis and formula semantic analysis. According to the characteristics of Chinese characters, this paper presents a method for automatic extraction of mathematical formula based on circular projection statistics. This method firstly collects key information through projection, and then extracts the potential line. Finally, mathematical formulas are extracted using a series of constraint conditions. The experimental results show that the method proposed in this work offers correctness of the results at very low computational costs.

Keywords: Mathematical Formula; Automatic Positioning; Circular Projection Statistics

基于循环投影统计的数学公式自动定位方法*

彭晓阳[1]，毛建频[2]

[1]邵阳学院经济与管理系，邵阳
[2]抚州职业技术学院，抚州
Email: ppakaka@qq.com

收稿日期：2013 年 4 月 13 日；修回日期：2013 年 4 月 21 日；录用日期：2013 年 5 月 20 日

摘　要：数学公式自动识别的第一步就是数学公式定位，只有从文档图像里正确定位出公式，后续的步骤如公式符号识别、公式版面分析、公式语义分析才能进行。本文根据中文文字特性，设计了一种基于循环投影统计的数学公式定位方法，该方法首先通过投影来统计关键信息，然后提取出可疑行，最后通过一系列条件进行可疑行的确认。实验结果表明本文提出的方法在计算成本非常低的前提下能保证结果的正确性。

关键词：数学公式；自动定位；循环投影统计

1. 引言

数学公式大量存在于各类科技文献之中，特别是在许多重要的文献中，由数学公式构成的科技准则常常占据着文献的核心地位。目前主流的 OCR 系统在处理文本方面已经具备很高的精确度和时效，但在处理数学公式方面还不尽如人意，当人们想要验证或想要重新使用数学公式时，只能借助于专门的数学排版工具或数学计算工具依据其规则重新输入，无法解决手动输入的低效率以及实现公式的自动化输入问题。

数学公式自动识别可以分为四个步骤：数学公式自动定位、数学公式符号识别、数学公式分析、公式

*基金项目：国家自然科学基金资助项目(61072121,61271382)；湖南省自然科学基金资助项目(12JJ2035)；江西省教育厅资助科研项目(GJJ11665)；湖南大学中央高校基本科研业务费资助项目。

分析结果输出[1]。数学公式自动定位作为自动识别的关键步骤之一，其定位的正确与否直接影响着识别的正确率。

对于将公式与文本分离，大致有两种方案：一种是排除法，即先对包含公式的文本进行识别，对于不能识别的部分就当成是数学公式抽取出来；另一种是根据公式本身以及排版的特性，对公式区域进行定位抽取[2,3]。对于第二种方法，基本上都是利用先验的知识进行处理，国外的做法有很多，但是经过对国外方法的研究，发现国外的方法较适用于外文，对中文并不是很适用，所以本文根据中文文字特性[4]，设计了一种基于循环投影统计[5,6]的方法进行数学公式的定位。

2. 预处理

从文档图像内正式定位数学公式之前，需要完成一些前期处理工作，这里主要是指各种预处理操作，通过预处理可以去除图像中无关的数据信息、保留有用的信息、极大地减少处理的信息量。本文采用的图像预处理过程主要包括二值化、噪声处理、图像的倾斜校正三个步骤，如图 1 所示。

2.1. 二值化

二值化是图像分割的一种方法。在二值化图像的时候把大于某个临界灰度值的像素灰度设为灰度极大值，把小于这个值的像素灰度设为灰度极小值，从而实现二值化。根据阈值选取的不同，二值化的算法分为固定阈值和自适应阈值[7]。比较常用的二值化方法则有：双峰法、P 参数法、迭代法和 OTSU 法[8]等。本文采用了 OTSU 法，即最大类间方差法，它是一种自适应的全局阈值选择法，此法既考虑到处理速度又保证了二值化的效果，能最大限度的使目标和背景相分离。其实现步骤如下：

1) 设 T 为前景与背景的分割阈值；

2) 计算出前景点数占图像的比例、平均灰度和背景点数占图像的比例、平均灰度。分别记为 W_0 和 U_0，W_1 和 U_1；

3) 设图像的总平均灰度为：$u = W_0 * U_0 + W_1 * U_1$；

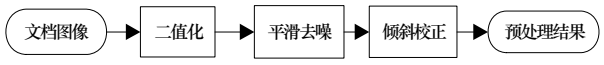

Figure 1. Pretreatment process
图 1. 预处理流程

4) 从最小灰度值到最大灰度值遍历 T，当 T 使得值 $g = W_0 * (U_0 - u)^2 + W_1 * (U_1 - u)^2$ 最大时，对应的 T 值则为分割的最佳阈值。

2.2. 平滑去噪处理

噪声的存在极大干扰了图像的信息，去掉噪声成分的过程叫图像平滑。一个较好的平滑方法应该既可以消掉噪声影响，又不会使图像的边缘轮廓和线条变模糊。图像平滑处理方法有空间域法和频域法两大类。在此，平滑去噪采用邻域平均法，它是一种空间域处理方法，基本原理为：设输入图像为 $f(x,y)$，则用邻域平均法得到的图像为 $g(x,y)$：

$$g(x,y) = \frac{1}{M} \sum_{(m,n) \in S} f(m,n) \tag{1}$$

式中 x，y 取值为 $0,1,\cdots,N-1$；S 为(x,y)点邻域中点坐标的集合；M 为集合 S 内坐标点总数。

在本文中，邻域的取法是以 $\sqrt{2}$ 为半径构成中心点 (x,y) 的邻域，选择在圆边界上的点和圆内的点为 S 的集合，即 $g(x,y) = \frac{1}{9}\sum_i f_i(m,n), i = 1,2,\cdots,9$。另外，为了减少因完全平均化而使图像边缘模糊的现象，本文还规定了当一些点和它邻域内点的灰度平均值差不大于规定的阈值 T 时，就仍保留其原灰度值不变，其中 T 是噪声成分标准差的常数倍，实际值由实验决定。

2.3. 图像的倾斜校正

由于本文的定位方法要用到投影操作，这种操作对倾斜比较敏感，故需要对图像对倾斜校正的操作。图像的倾斜校正一般分为手动校正和自动校正。本文中采用的方式是手工校正。

3. 数学公式定位

图像经过预处理后，便是如何抽取公式了，数学公式定位又包含独立公式定位和内嵌公式定位两种方式，内嵌公式是指和文字夹杂在同一行的公式，其中，独立公式和普通文字行的区别较大，可以根据版面先验知识直接定位，内嵌公式的定位则需要进一步确认。本文数学公式的自动定位的实现分为如下三步：

一、统计关键信息：首先进行第一次行投影，统计得出普通行宽度、普通行间段宽度、普通行密度，

并记录每行的位置。然后对刚刚行投影得出的单独行进行列投影，根据统计得出每行的普通文字行数字个数、汉字宽度、和汉字间距等信息。

二、提取可疑行：进行行投影，根据一系列条件抽取出可疑的行，待最后一步确认。

三、确认可疑行：即对步骤二提取出的可疑行进行一系列的确认。

3.1. 统计关键信息

3.1.1. 行投影

设文本图像大小为 $G_x \times G_y$，文本的二值图像为 $f(i,j)$，其中，i、j 分别为像素的行、列坐标。那么，定义 $f(i,j)$ 在第 i 行上的投影函数为：

$$g(i)=\sum_{j=1}^{G_y} f(i,j), i=1,2,\cdots,G_x \tag{2}$$

可以看出，若该行为行间隔，则 $g(i)$ 为 0，那么首先对 $g(i)$ 二值化得出二值序列 $g_1 g_2 \cdots g_{M_x}$。然后通过以下几个步骤来统计分析这次行投影数据。

1) 根据二值序列 $g_1 g_2 \cdots g_{M_x}$ 得出所有行段和行间段宽度，并记录下每行的位置信息，存入一个二维数组 $R[\text{rowNumber}]=[\text{rowHead}，\text{rowTail}]$。

2) 对宽度进行统计分析，将出现次数最多的行段距离作为普通行宽度，记为 ROWWIDTH，也用同样方法得出普通行间段宽度，记为 SPACEWIDTH。

3) 计算各非间隔行的行密度：

$$\text{density}=n/(\text{width}\times M_y) \tag{3}$$

式中 n 为行段的黑色像素总数，width 为行段宽度。然后根据普通行宽度 ROWWIDTH，即抽取所有行宽趋近 ROWWIDTH 的行，统计所有普通行密度，求出其平均值，记为 AVRDESTINY。

3.1.2. 列切分

遍历行切分得到的所有行，如果行段满足

1) 行段密度大于平均密度 AVRDESTINY；

2) 行段宽度非常趋近于 ROWWIDTH，则抽取出该出行。

设抽取得到的图像为 $G_x \times R_y$，定义“抽取行”在 j 列的投影函数为

$$c(j)=\sum_{i=1}^{G_x} f(i,j), j=1,2,\cdots,R_y \tag{4}$$

则列切分的方法如下：

1) 寻找满足 $c(j)=0$ 且 $c(j+1)\geq 1$ 的点，作为单独字符起始位置 j_{head}；继续寻找满足 $c(j-1)\geq 0$ 且 $c(j)=0$ 的点作为单独字符的终止位置 j_{tail}。

2) 按(1)方法找出所有字符，统计字符数量，然后根据最常出现的数量作为普通文字行数字个数，记为 RWORDNUM。

3) 算出所有的 $w_1=j_{\text{tail}}-j_{\text{head}}$，统计最常见的 w_1 作为汉字宽度，记为 WORDWIDTH。

4) 算出所有的 $w_2=j_{\text{head+1}}-j_{\text{tail}}$，统计最常见的 w_2 作为汉字间距，记为 WORDSPACE。

3.2. 提取可疑行

在统计完基本信息后。则可以进行可疑行的提取了。判断可疑行的条件主要有以下三个：

a) 当前行密度是否小于普通行密度 AVRDESTINY。

b) 当前行宽度和普通行宽度差别较大。

c) 当前行上下间隔比行平均间隔略大。

另外，还要考虑对含有上下标公式的处理。经分析，如果投影后上小标为独立的一行，那么通常情况下其与公式行的距离要远小于普通行间段宽度 SPACEWIDTH。所以本文采取的方法是发现这种情况则判断离它最近的行是否为公式行(即是否满足上述三个条件)，是则把它与该行合并后直接抽取。如图 2 所示是提取可疑行的程序流程图。

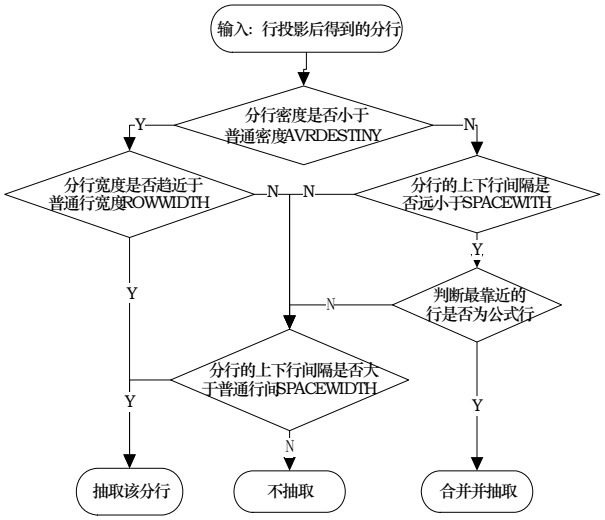

Figure 2. The flow diagram of extract the suspicious line
图 2. 提取可疑行的程序流程图

经过上一步骤，就完成了可疑行与正常文本行的分离工作。但是很明显并不能认为这些可疑行就是公式区域了，如前所述数学公式包含独立公式和内嵌公式两种，所以对于夹杂在文本行中的内嵌数学公式，还是得进一步确认。

3.3. 公式区域确认

在确认步骤中，充分利用了汉字是方块字的特点。首先对可疑行进行列投影，对其中的字符进行统计分析，最后结合前文中收集的汉字信息来确认公式区域。对每一单独可疑行的具体操作流程如下：

1) 设可疑行为 $G_x \times R_y$，进行列投影，统计得出该行普通字符宽度 wordWidth、普通字符间距 wordSpace、总字符个数 wordNum。并定义单个字符投影位置为 $w_i = \left[w_{起始}, w_{终止}\right]$，$i = 1, 2, \cdots, R_{\text{wordNum}}$，则得到字符序列 $w_1 w_2 \cdots w_{\text{wordNum}}$。

2) 当 $\text{wordNum} < \text{RWORDNUM} - \alpha$ 或 $\text{wordNum} > \text{RWORDNUM} + \beta$ 时，则判定为独立公式行。否则继续。(α，β 为经验值)。

3) 遍历字符序列 $w_1 w_2 \cdots w_{\text{wordNum}}$，若有连续四个以下字符满足以下三个条件，则可判定为内嵌数学公式，则把满足条件的最大子序列抽取出来即可。

a) 该字符宽度不趋近于 WORDWIDTH；

b) 该字符左右空白的宽度不趋近于 WORDSPACE；

c) 查找该字符其后紧接着的字符以及空白宽度，并也满足(a)、(b)两个条件。

最终遍历完所有可以行后，就排除了图像中的文本信息，只剩下含有数学公式的区域了。

4. 实验结果及分析

本文所述的预处理和定位算法都采用 C 语言实现，实验环境是 Visual Studio 2012，共对 100 个包含数学公式的文献截图进行数学公式定位。经过实验统计，对独立的公式行，不论在文档内含有单个独立公式行还是多个独立的公式行，本文的定位方法能够达到较好的效果，定位正确率可达 90%左右。而对于内嵌数学公式的定位要比独立公式行的定位困难，其定位率也明显下降，就实验所用数据来看，正确率可达 70%左右。

如图 3 所示是公式定位前的文档图像，如图 4 所示是公式定位后的图像。

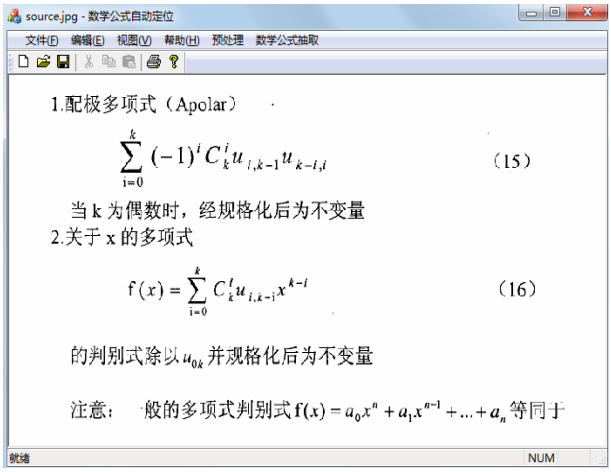

Figure 3. The picture before formula location
图 3. 公式定位前的文档图像

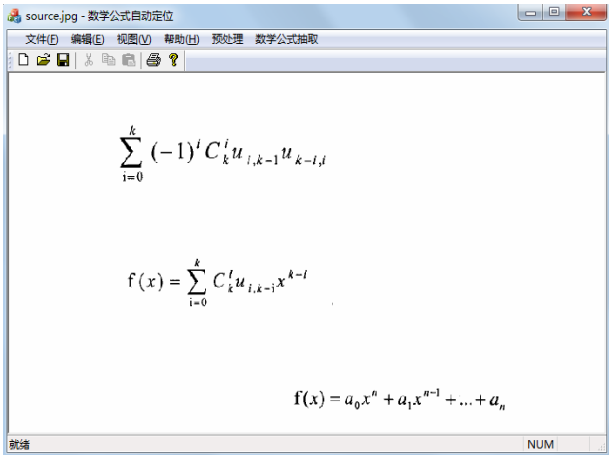

Figure 4. The picture after formula location
图 4. 公式定位后的图像

5. 结论

本文根据汉字是方块字的特点，设计了基于循环投影统计的数学公式定位方法，记录出汉字区域与公式区域的差别，包括字符宽度、间距和密度等，并以这些差别作为公式区域确认时的参考，从而定位出中文文献中的独立数学公式以及内嵌数学公式，并最终通过 C 语言编程验证，证明本文的方法是可行的。

本文工作还有一些不足之处：

1) 输入的文档图像有时候不仅包含数学公式和普通文本，还有可能包含图表、图片。本文只处理了包含数学公式与文本的图像，对于包含图表、图片等的混合图像还需要进一步的研究。

2) 本文利用汉字特点设计出的公式定位方法对于矩阵等复杂的公式并不能适用，如何更好地利用汉字特点，也还需要进一步地研究。

3) 图像中存在的公式会对统计关键信息步骤造成偏差，并且如果公式较多，甚至会支配以上关键信息的取值，这也是需要进一步研究的地方。

参考文献 (References)

[1] 程进. 基本数学公式识别技术的研究[D]. 沈阳工业大学, 2004.

[2] 陈峰, 郑春光. 印刷体文档中的数学公式识别方法综述[J]. 信息技术, 2009, 3: 15-23.

[3] K.-F. Chan, D.-Y. Yeung. Mathematical expression recognition: A survey. International Journal of Oil Document Analysis and Recognition, 2000, 3(1): 3-15.

[4] 丁晓青. 汉字识别研究的回顾[J]. 电子学报, 2002, 30(9): 1364-1368.

[5] 章毓晋. 图象分割[M]. 北京: 科学出版社, 2001.

[6] 刘立波. 图像分割方法探讨[J]. 宁夏农学院学报, 2001, 22(4): 51-56.

[7] 吴冰, 秦志远. 自动确定图像二值化最佳阈值的新方法[J]. 测绘学院报, 2001, 18(4): 283-286.

[8] 张洪刚, 陈光, 郭军. 图像处理与识别[M]. 北京: 北京邮电大学出版社, 2006.

